CIENCIA DE DATOS, *ANALYTICS*, *MACHINE LEARNING*, INTELIGENCIA ARTIFICIAL, DATA *WAREHOUSE* Y *BUSSINESS INTELLIGENCE* PARA SUBSUELO

Data analytics en *tight gas*

Por ***Diego Gallart y Andrés López Gibson (Y-TEC)***

En este trabajo se presenta la aplicación de técnicas de *Data Science* en las diferentes especialidades para el procesamiento y la interpretación de sus datos en bruto, con aplicación concreta al *tight gas.*

Históricamente la información de los activos se almacena en diferentes formatos, incluso no digitalizados. En el trabajo que presentamos, el volumen y la variedad hicieron que sea difícil su análisis integrado en una primera instancia. En la actualidad, con avances en el tratamiento de los datos, su estandarización, su integración continua y las herramientas para explotar su valor es posible y necesario avanzar sobre proyectos de integración en repositorios disponibles para el análisis en conjunto de los especialistas del negocio y los especialistas de data analytics.

El trabajo incluye la selección de datos y sus fuentes, la estandarización, la integración y el aseguramiento de la calidad para su posterior modelado, análisis y generación de recomendaciones.

Implica trabajar con datos crudos e interpretados disponibles en diferentes repositorios de la empresa, incluidas las siguientes etapas: preparación de datos, adición de variables calculadas basadas en el conocimiento del dominio y análisis multivariado.

Como resultado de este trabajo se crea un repositorio de datos estandarizado e integrado que podría actualizarse posteriormente cuando haya nuevos datos disponibles.

La explotación de los datos se desarrolla con técnicas de Analytics, machine learning y herramientas de visualización, buscando generar recomendaciones para futuras terminaciones de pozos y optimización de producción, junto con una mejor comprensión de las relaciones entre petrofísica, producción y estimulaciones.

En este trabajo se presenta la aplicación de técnicas de *Data Science* en las diferentes especialidades para el procesamiento y la interpretación de sus datos en bruto. Se exploran los datos en busca de relaciones inesperadas entre variables, como geomecánica, *cuttings*, PLT, especificaciones sísmicas y de fractura, con el objetivo de resaltar los aspectos particulares que optimizan el desarrollo de un campo de *gas tight*. En el futuro cercano, el procedimiento podría extrapolarse a otros activos con formaciones análogas.

Desarrollo técnico del trabajo

Datos

En este primer trabajo se integran datos de diferentes especialidades disponibles al momento de su ejecución. Como condición inicial se requiere que el pozo tenga PLTs. Luego, datos de estimulación crudos e interpretados. Finalmente, se incorporan datos de Geología, Geofísica, Geomecánica y Petrofísica. Para poder sincronizar el análisis en el tiempo para todos los pozos se incorpora la producción histórica diaria (Figura 1).

Se dispone en el set de más de 50 pozos. Con un promedio de 12 etapas de fractura por pozo. Lo que da más de 600 observaciones para el análisis.

El nivel de detalle de los datos se enfocó en la etapa de fractura, porque en la granularidad eran compatibles todos los orígenes de datos, fue factible, por un lado, agregar los datos de perfiles en profundidad y, por otro lado, distribuir la producción en el tiempo del pozo a cada etapa sin perder el nivel de detalle del análisis.

Figura 1. **Repositorio de datos.**

Ensayos de producción (PLT)

La producción se distribuyó en las etapas de fractura según la interpretación de los PLTs. En algunos casos dos o más etapas de fractura se encontraban agrupadas, principalmente porque el perfil de medición no alcanzaba la profundidad total para poder separar los caudales producidos. Para esos casos se separaron los datos con diferentes criterios. Inicialmente, se buscaron PLTs anteriores y posteriores con mayor nivel de detalle, de existir, la producción se estimó interpolando las mediciones de los distintos PLTs. En caso de no contar con referencias de otros PLTs cercanos, se aplicó una regla por espesor de las fracturas involucradas para distribuir la producción. Esto se realizó para el 10% de los PLTs.

En el caso de la temperatura y la presión, cuando se detectaron faltantes, se infirieron valores entrenando modelos de regresión para predecir los datos necesarios en base a los existentes en el PLT particular y los restantes PLTs del mismo pozo (problema fondo de pozo). Esto se realizó para el 25% de los PLTs (Figura 2).

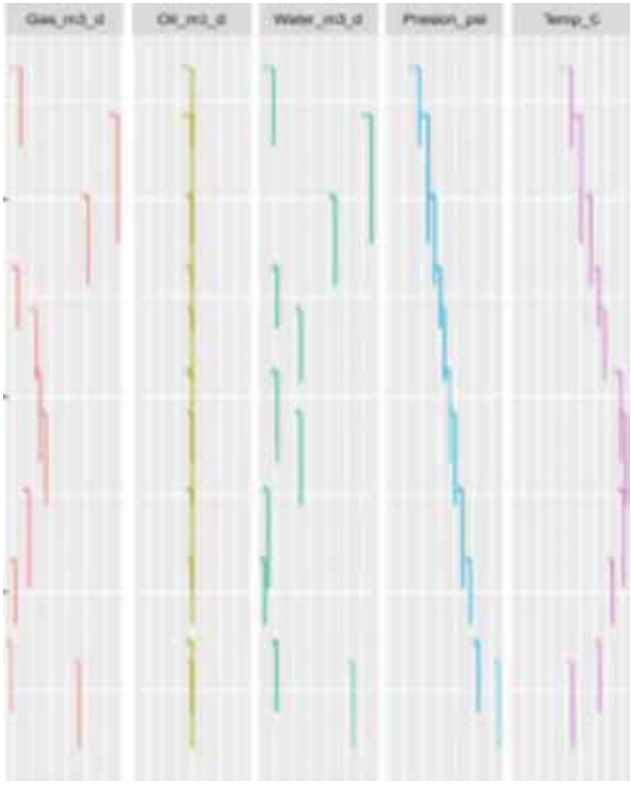


Figura 2. **Ejemplo de datos de PLT.**

Producción diaria histórica

Para poder sincronizar el análisis en el tiempo se incorpora la producción histórica diaria de cada pozo. Se llevó la granularidad de análisis al nivel diario, ya que se buscó darle importancia al período de apertura, limpieza y *flowback* de cada etapa, así como también se intentó capturar la mayor cantidad de información de apertura y cierre de orificios que se dispusiera durante la historia de producción. Por último, se incorporó la declinación de reservas al Q3 2019 de cada pozo para extrapolar el análisis hasta EUR. Este *forecast* se realiza con el mejor

ajuste de un declino inicialmente hiperbólico y luego exponencial hasta el fin de vida útil del pozo.

Al integrar la producción diaria con la información de los PLTs y distribuyéndola por etapa de fractura en el tiempo, logramos tener las producciones diarias y acumuladas a 30, 60, 90, 180, 360, 1000 y EUR días por etapa desde la fecha de fractura.

En los casos de datos faltantes por un período de tiempo, normalmente ocasionado por un error de medición de los sensores o por falta de medición, se imputaron los valores interpolando entre las mediciones con modelos de regresión. Los datos calculados más los datos interpolados representan un 8% del set de datos.

Estimulación

Los datos de estimulación incorporados al estudio son datos crudos y datos interpretados. Un total de más de 60 variables que comprenden datos de tipo y fecha de fracturas, *minifrac*, *dfit*, tipos y volúmenes de fluidos y agentes de sostén, sus caudales y concentraciones obtenidos de cartas de fractura y las interpretaciones de tamaño de fractura estimadas por el especialista.

Geología y geofísica

Con respecto a geología se incorporan los topes de las superficies geológicas, interpretados por los especialistas, para cada pozo. Esa información se integra con los topes y bases de las fracturas para obtener las superficies contactadas por cada fractura. Se registran las cinco superficies de mayor contacto por etapa de fractura y las proporciones de fractura que contacta cada superficie. Para la especialidad de geofísica, se incorpora mediante archivos planos y segy las fallas, sus coordenadas y atributos calculados a partir de la inversión sísmica. Para cada etapa de la fractura se calculó su centroide y luego se registraron las propiedades cercanas, a menos de 500 m, y sus distancias.

Petrofísica y geomecánica

Los datos de petrofísica y geomecánica se dispusieron en archivos.las (Log ASCII Standard). Estos sets se procesaron tomando los segmentos correspondientes a los tope y base de las etapas de fractura, distinguiendo zonas RES y PAY y calculando medidas de resumen, como el promedio, mediana, máximo, mínimo y desvío estándar. También, se calcularon las proporciones de RES y PAY para cada etapa de la fractura.

Las variables de petrofísica incorporadas al estudio comprenden datos de control geológico, como cromatografía y litología en superficie llevada a profundidad; datos de perfiles a pozo abierto, como gamma ray, resistividades y sónicos y; perfiles interpretados, como porosidad, permeabilidad y saturación de agua. Los datos de geomecánica incorporados al estudio comprenden datos de eventos de pozo y de control geológico llevados a profundidad y perfiles interpretados de propiedades mecánicas, elásticas y de resistencia, estado de esfuerzos y gradientes de fractura.

Variables calculadas y categóricas

En base a recomendaciones de los expertos de cada especialidad se calcularon fórmulas (*Feature Engineering*) que sirven de resumen de varias variables.

Las variables integradas al estudio comprenden medidas agregables, cocientes de rendimientos, tratamiento de superposiciones entre las fracturas, proporciones de contacto de las formaciones geológicas y variables que resumen la calidad de la etapa y se correlacionan con su rendimiento, como PAY contratado y HCPV (*HidroCarbon Pore Volume*). Las variables categóricas de texto fueron convertidas a un ID numérico con el fin de analizar linealidad y poder aplicar filtros mayor/menor.

Variables objetivo

Se calcularon diferentes variables objetivo para su posterior análisis. Inicialmente, la producción acumulada por fractura (etapa de fractura) a 30, 60,9 0, 180, 360, 1000 días y EUR. Luego se decidió quitar el efecto del espesor de fractura y, por eso dividir la producción acumulada por los metros de PAY contactado. Para considerar la superposición de fracturas, se creó otra variable objetivo a producción acumulada EUR normalizando por el espesor asignado. También se calcularon estas variables

de rendimiento por HCPV.

Con el objetivo de entrenar modelos de clasificación se discretizaron los valores convirtiéndolos en FLAGS (valores 0 o 1) que indican si el rendimiento corresponde a los cuantiles 25, 50 o 75.

Finalmente, incorporamos el valor económico al objetivo y calculamos el costo y beneficio esperados por fractura. Para esto se estimaron los costos por cantidades y tipos de fluidos, agente de sostén y fracturas por pozo. Para el cálculo del beneficio se consideró el valor de la producción con los precios estimados del hidrocarburo en el tiempo.

Para estos objetivos económicos se utilizaron técnicas de Valor Presente y Valor Futuro para el cálculo del Valor Actual Neto (VAN e IVAN).

Análisis multivariado

Dentro del análisis multivariado se buscó medir y visualizar relaciones entre las variables integradas para responder preguntas del negocio y avanzar hacia la construcción de los modelos predictivos.

Se analizaron relaciones de las etapas de fractura con:

- Posicionamiento en la estructura geológica, formación, saturación de agua y producción.
- Superposición entre etapas y rendimientos.
- Rendimientos por PAY contactado, espesor de etapa y NTG (Net To Gross).
- Rendimientos *versus* tiempo de residencia de fluido.
- Rendimientos considerando arenamiento.
- Rendimientos *versus* máximas concentraciones de arena en fondo y bombeada.
- Rendimientos por cantidad de fluidos y tamaño de fractura.

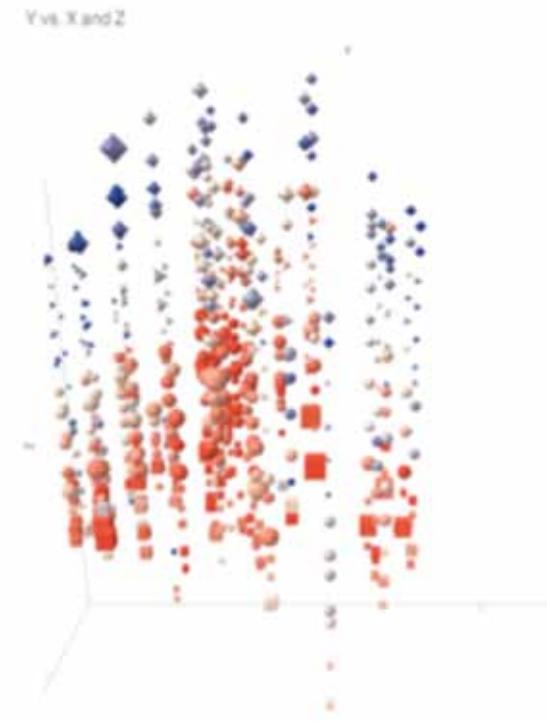


Figura 3. **Etapas 3D y saturación de agua en PAY contactado.**

- Rendimientos por diseño de estimulación (agente de sostén/PAY).
- Rendimientos por número de *clusters*.

Modelos predictores

En este trabajo de construyeron una gran variedad de modelos predictores con distintos objetivos de interés según lo detallado en la construcción de variables objetivo. Para cada modelo el proceso incluyó un análisis y selección de variables y luego la selección del modelo más adecuado (regresión o clasificación) priorizando primeramente su interpretabilidad.

Selección de variables

Para cada modelo, para cada variable objetivo es necesario realizar una selección de variables.

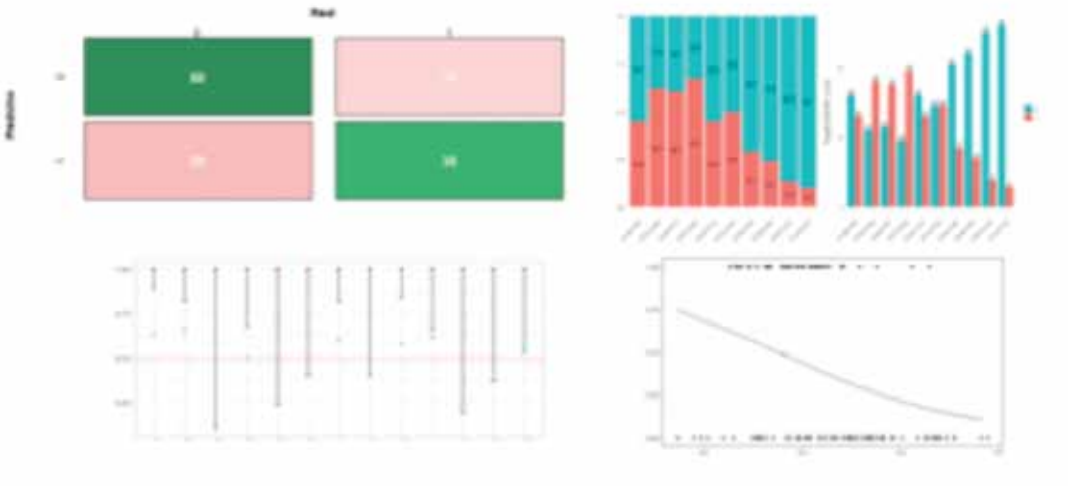


Figura 4. **Análisis de modelos económicos.**

Para los modelos de clasificación se seleccionan variables mediante algoritmos de selección automática de variables. Ese resultado es analizado por los especialistas y se seleccionan dentro de las variables con mayor significancia, las que sean de mayor interés para las especialidades. Con este conjunto de variables preseleccionadas se entrena un modelo con el set de entrenamiento reservado para este fin.

Modelos económicos

Como modelos económicos se crearon predictores para IVAN para 360, 1000 días y EUR. Se discretizaron estas variables con la regla de que si costo/beneficio es menor a 1 el beneficio no llegó a cubrir el costo y se le asigna un FLAG=0 (falso). Se asigna FLAG=1 en caso contrario, el conjunto es de las etapas de fractura cuyo beneficio fue mayor al costo.

Para evaluar este tipo de modelo se utilizaron matrices de confusión sobre un conjunto de datos de validación reservado y se compararon los F1 Scores.

El F1 Score es la media armónica entre exactitud y exhaustividad y servirá para comparar entre modelos. Para cada modelo mediante el análisis de la curva ROC se calcularon los umbrales de discriminación que optimizaban el F1 Score.

Mediante visualizaciones se explicó el efecto de cada variable seleccionada y se analizaron los errores de cada tipo en los modelos.

Para la construcción de los modelos a 1000 días y EUR se consideraron los resultados de los modelos previos en el tiempo y se incorporó su resultado al nuevo modelo creando así un ensamble de modelos (Figura 4).

Modelos de producción

Para la predicción de producción se construyeron modelos de regresión para producción acumulada a 360, 1000 días y EUR midiendo su precisión por RMSE y explicando los efectos de las variables seleccionadas por el estudio de su aporte a las predicciones particulares (Figura 5).

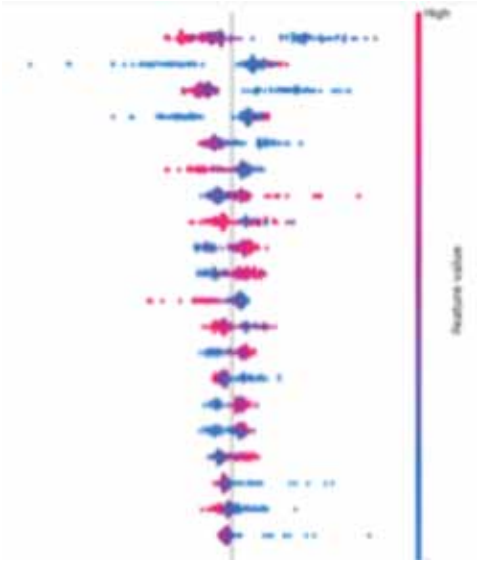


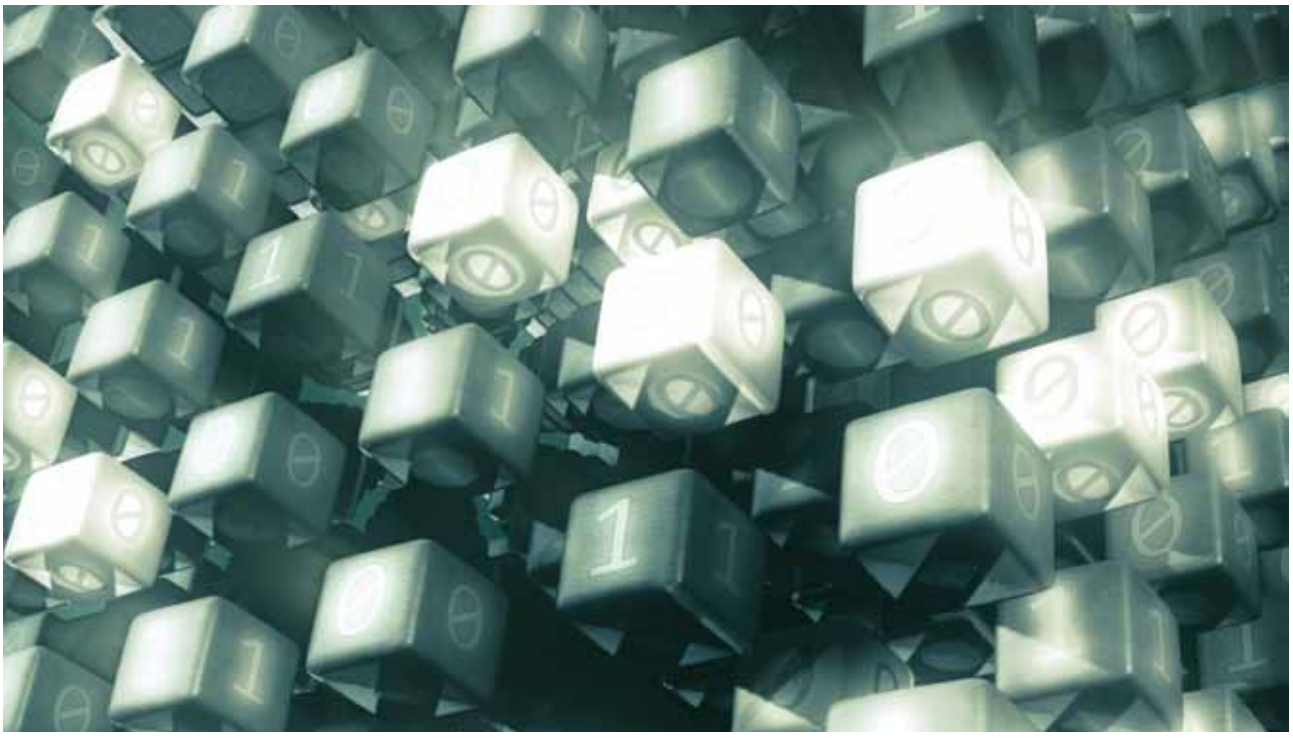
Figura 5. **Análisis modelos de producción.**

Resultados obtenidos

Durante este trabajo se lograron identificar los orígenes de datos y su perfilado para su análisis integral. Ese proceso, normalmente llamado *data pipeline,* será implementado con un conjunto de normativas de puntos de entrega y formatos para datos futuros. De esta forma se podrá disponer de los datos actualizados en un repositorio único integrado y podrán ser explotados con herramientas de análisis por los especialistas.

Con los datos integrados en un repositorio único se configuraron herramientas de visualización que permitieron a los especialistas tener acceso a los datos y compartir sesiones de análisis con los científicos de datos. De esta forma se pudo poner foco en el análisis multivariado a los aspectos que los especialistas hallaban de interés.

Se desarrollaron primeras versiones de algoritmos que permiten ajustar la granularidad de los datos al nivel de estudio (etapa de fractura) que pueden ser perfeccionados con hallazgos resultado de este trabajo.



El equipo definió variables económicas de cálculo simple como objetivo de los modelos. Para ello consideró el costo/beneficio a valor actual neto para determinar si una etapa de fractura era rentable o no.

Para el modelo de predicción económico a un año (360 días) se observó una precisión en el set de prueba del 74% y que las variables con más impacto en la predicción fueron la saturación de agua, el HCPV, la coordenada X, la cantidad de agente de sostén por metro, la porosidad y la coordenada Y de la fractura. Las predicciones realizadas con este modelo fueron utilizadas como input del modelo de 3 años (1000 días) y por eso se consideran de importancia también para ese modelo.

El modelo de predicción económico a 3 años (1000 días) resultó con una precisión menor, del 70% en el set de test, y las variables importantes para este modelo fueron, adicionales a las del modelo de 1 año: la permeabilidad, la resistividad, el RHOB, la cantidad de PAY contactado, la relación PAY/Reservorio contactadas y la cantidad de fluidos totales por HCPV.

Para el modelo de predicción económico a vida útil del pozo (EUR) se obtuvo una precisión del 82% en el set de prueba. Las variables de mayor importancia, adicionales a las del modelo de 1000 días, fueron la resistividad y la permeabilidad del NO PAY contactado.

El impacto en el negocio abarca desde intangibles, como el ordenamiento de la información disponible, pasando por tangibles como hallazgos de datos que no se estaban utilizando en las bases relevadas. Los hallazgos motivaron también oportunidades de optimización que serán puestas a prueba a través de pilotos en campo.

Conclusiones

Las tareas de búsqueda e integración de datos fueron las que más tiempo consumieron dada la diversidad de la información y los formatos existentes. Durante los años no se ha establecido un formato único de entrega de resultados de los estudios y esto ha llevado a encontrar datos en formatos que hacen difícil la automatización de la carga, incluso con la única opción de carga manual.

Durante el transcurso del trabajo se realizaron reuniones de consulta y presentación de resultados con el equipo. Esto fue clave para comprender muchos aspectos de las especialidades que permitieron refinar el estudio. Los especialistas dieron recomendaciones sobre el tratamiento de las variables, sus rangos de valores y sus umbrales de discriminación. Esto permitió integrar las variables al estudio maximizando su aporte.

Como resultado de las charlas con los especialistas se concluyó que era necesario mejorar las variables económicas considerando particularidades con modelos más complejos, por ejemplo, costo marginal, tipo de estimulación, calidad de arena y fluidos utilizados. Es necesario aclarar que este aspecto como otros el trabajo es un primer acercamiento y permitirá refinar los procesos con las lecciones aprendidas.

Por ello fue fundamental trabajar con una comunicación fluida con el equipo de especialistas integrándolos al trabajo en continuo. Como lección aprendida, más allá de lo técnico, para este tipo de trabajos es fundamental formar un equipo multidisciplinario que integre el rol de científico de datos al equipo y reservar el aporte de cada especialista al estudio.

Comprendimos la criticidad que tienen las buenas prácticas en la toma y el resguardo de datos en forma temprana y su impacto en el negocio.

Por último, se considera que este trabajo podría ser extendido a otros yacimientos similares que se beneficien con los modelos y los flujos de trabajo generados. Al haber una extensión areal importante de la formación, esto contribuiría a incrementar el beneficio mutuo.